Machine Learning Modelling to Predict Lung Cancer Stages from CT Scan Image

Sabina Zaman[1*], Mamun Ahmed[2], Mohammad Asaduzzaman Khan[2], Tamanna Tajrin[2]

[1,2]Department of Computer Science and Engineering, Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh

*sabinazaman204@gmail.com

Abstract: Lung cancer is considered the most common cancer for men and the third common for women. According to the world health organization report near about 1.76 million people died from lung cancer in 2018. Among various computer-aided diagnosis systems processing and analysing CT scan images to detect cancer from images of nodule has become popular in this age. After the implementation of several image processing steps, four (4) significant features- Area, Eccentricity, Diameter, and Perimeter have been extracted. Not only from online CT images of the lung but also using real-life data, a custom database has been prepared. As it is a self-made database, class labels have been determined according to standard rules for stage labelling, so the number of clusters has been verified using the K-valid algorithm. For classification purposes of cancer nodule staging, various machine learning algorithms have been implemented. The comparisons of accuracy and other measures of the classifiers have been implemented to rate and to choose the best classifier for this subject. It is observed that the overall accuracy of each machine learning algorithm has been improved after implementing new approaches to image processing. Unlike other approaches of binary class prediction and implementation of a single algorithm for the task, here we have tried to predict stages and a comparison among the three most traditional machine learning algorithms has been demonstrated.

Keywords: *Lung cancer, CT scan images, Image processing, Machine learning algorithm*

1. Introduction

When any cell or cells of the lung grow and multiply abnormally without any control, it results in a tumor. If the tumor starts to spread around nearby tissues, lymph nodes, or distant parts through blood or lymphatic circulatory system, it is called cancer or cancerous tumor in the lung. Confiding on tumor size and nature of spread cancer has been divided into four stages. [1]

Though lung cancer is a deadly disease, the survival rate increases with early diagnosis followed by proper treatment. It is stated from the National Statistics Office UK that; 15-19% of patients of stage IV patients have 1-year survival rate while it increases to 81-85% for stage I patients [2]. There are many techniques to detect lung cancer stages like X-ray or (CXR), CT or Computed Tomography image, liquid biopsies, using antibodies, circulating microRNAs and so on [3]. But from CXR images we don't get detailed aspects and differences between images of different time are tough to notice while low dose CT is more effective to diagnose cancer as it combines cross-sectional X-ray images along with the computer. And pathological tests like liquid biopsies or antibodies are more complex, costly and time consuming. Recently various computer-aided systems depended on image processing and machine learning algorithms have been introduced to reduce time and complexity to segment and to classify cancer nodules.

2. Related Works

For a computer-aided diagnosis system, mostly CT images are used. G.N. Hounsfield who worked at the Central Research Laboratories of EMI Ltd constructed a CT imaging system. Various techniques are used to process images in several steps like pre-processing, noise removing, or performing segmentation to extract valid features. M. A. Gajdhane used a median filter to remove noise, Gabor filter for enhancement purpose, and Marker-controlled watershed segmentation [5].

To extract features from the lung cancer nodule, mean and median filter both are used to reduce noise as part of preprocessing. Marker-controlled watershed and Otsu's Thresholding are implemented for segmentation [6].

Table 1 Overview of percentage of deaths due to various cancers [3]

Cancer Site	No. Of New Cases (% Of All Sites)	No. Of Deaths (% Of All Sites)
Lung	2,093,876 (11.6)	1,761,007 (18.4)
Breast	2,088,849 (11.6)	626,679 (6.6)
Prostate	1,276,106 (7.1)	358,989 (3.8)
Colon	1,096,601 (6.1)	551,269 (5.8)
Non-melanoma of skin	1,042,056 (5.8)	65,155 (0.7)
Stomach	1,033,701 (5.7)	782,685 (8.2)
Liver	841,080 (4.7)	781,631 (8.2)
Rectum	704,376 (3.9)	310,394 (3.2)
Esophagus	572,034 (3.2)	508,585 (5.3)
Cervix uteri	569,847 (3.2)	311,365 (3.3)

A model was proposed by A. Amutha and R.S.D Wahidabanu where Active Contour Modeling was used to diagnose lung cancer[7] [8]. Auto enhancement as

preprocessing, Gabor filter and Fast Fourier transform (FFT) to enhance the image and Thresholding and Watershed segmentation as segmentation were used in another work. To extract feature, Binarization and Masking approaches were implemented, to extract an exact region following the region growing method, after threshold segmentation. N.A. Memon et. al [9].

To detect cancer nodule's position and staging of cancer, classification, and clustering are done by different machine learning algorithms which are categorized into supervised and unsupervised classification techniques.

An automated segmentation method based on a knowledge-based fuzzy system was first introduced by Brown, McNitt and N J Mankovich [10]. A hybrid fuzzy system for a scattered search to make segmentation of CT image was implanted later. Here evolutionary optimization was done to reduce population [11].

After proposing CNN with 2 hidden layers, [12] later combination of fuzzy logic and NN was introduced which provided better results than the cohesion of the Genetic algorithm and CNN [13]. Decision Tree, Feed Forward Neural Network were compared while features were extracted from the Low dose CT image [14].

J.Wang, R. Engelmann and Q. Librought in the idea of CNN while [15]T. Lin and C.-R. Yan exhibited that Fuzzy Logic and NN could give better results if they are merged into a single system[16].

3. Simulation Details

Here, DICOM formatted CT images which are available in The Cancer Imaging Archive, CMH Cumilla and Cumilla Central Hospital have been used. As pre-processing of dataset some image processing procedures have been accomplished like smoothing to remove noise, contrast correction, morphological operation, outlining, and region extraction, edge detection and segmentation operation. And then nodules are separated perfectly from the background using object measurement processes. After reviewing several papers [17], [18], we have considered four features which are Area, Eccentricity, Diameter, and Perimeter for classification purpose. After the feature extraction step has been performed, custom dataset has been prepared. Total 317 samples are used as training samples to train 3different networks and 140 samples are available for the testing purpose.

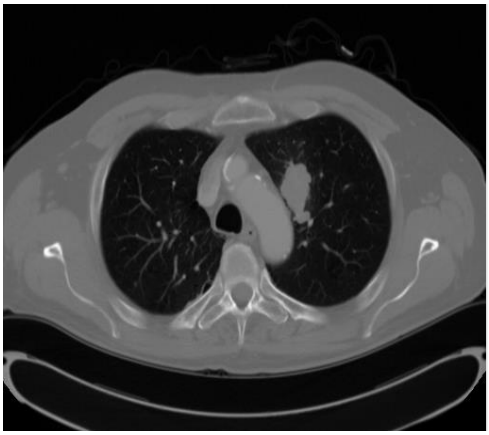

Fig.1.*Original Image.*

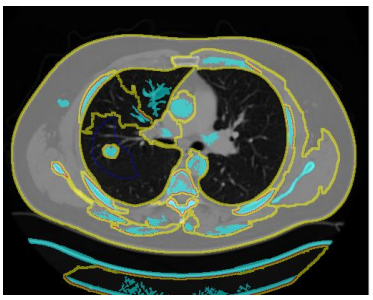

Fig.2. *Soft tissues are detected as cancer cell.*

Table 2 Sample Data

Area (pixel value)	Eccentricity (pixel value)	Diameter (pixel value)	Perimeter (pixel value)	**Stage**
1081	0.38661	37.099	132.35	1
1257	0.67322	40.006	207.57	2
1429	0.88657	42.655	190.59	2
138	0.65881	13.255	46.735	1
1658	0.60768	45.946	154.66	1
267	0.87578	18.438	77.426	1
418	0.54017	23.07	73.654	1
1010	0.80024	35.86	226.88	1
497	0.71477	25.156	87.094	1
1576	0.74324	44.795	184.02	2
13355	0.76448	130.4	618.9	4
15121	0.36916	138.75	581.76	4
366	0.86592	21.587	77.616	1
1510	0.68675	43.847	240.04	2
3316	0.59722	64.977	229.42	3

Table 3 Characteristics of Attributes

Sl	Attribute	Minimum Value	Maximum Value	Std. Deviation
1.	Area	82	18081	6019.828
2.	Eccentricity	.369	.903	.151
3.	Diameter	3.847	151.73	45.768
4.	Perimeter	46.735	767.58	214.194
5.	Stages	1	4	-

Various algorithm parameters like batch size of Decision Tree or gamma value of Support Vector Machine have been selected by fine tuning procedure.

3.1. K-valid Algorithm

As we have considered 4 clusters or stages of cancer, the number of clusters has been proved by the K-valid algorithm. For this purpose, K-means algorithm and Silhouette-Index or Elbow method are used together. K-means is a type of unsupervised machine learning algorithm to categorize data samples into K number of groups where usually K varies from 1 to 10.

Now, to determine the best value of K, the elbow method is used. Generally, it is determined visually from a graph where the sum of square distances and K values are plotted against one another. When the chart appears as an 'arm' and

the 'elbow' where the inflection starts, will be considered as the best value of K.

K-means algorithm-

1. K is chosen as number of clusters.
2. Choose randomly K number of samples from dataset as initial members or centre points of K classes
3. do
4. (re) assign each sample to a specific class calculating minimum Euclidean distance from each centre.
5. Update cluster centre averaging samples of the specific class.
6. While no change of cluster centre.

Elbow Method-

1. Generate Sum of Squared Error (SSE) for each K value
2. Draw a graph plotting cluster number K vs SSE.
3. The graph will look like and then the 'elbow' with x value of K is the best.

3.2. Naïve Bayes Classifier

This machine learning algorithm calculates class membership likelihoods for a given sample and is considered to bethe member of any class with maximum possibility. It is supposed that each attribute is independent of the other. Bayes theorem is used in Naïve Bayes Classifier. Suppose A is a sample with n number of features and Hy can be any sort of hypothesis depending on which A becomes a member of class X. In Bayesian terms A is known as 'evidence'. So now we have to measure Probability (Hy|A), the possibility of A to be belonged to class X from given attribute narration for Hypothesis Hy.

Probability (Hy|A) is posterior possibility of Hy conditioned on A. Probability(A|Hy) posterior possibility of A conditioned on Hy, Probability(Hy) prior possibility of Hy and Probability(A) prior possibility of A.

So, Probability (Hy|A) = $\frac{Probability\ (A|Hy)Probability(Hy)}{Probability(A)}$

Algorithm

1. Suppose Dt is full dataset with n number of attributes $Attributes_1$, $Attributes_2$, $Attributes_3$,…$Attributes_n$ and so any sample S = (s_1,s_2,s_3,…s_n)
2. If there are m number of classes like $Class_1$,$Class_2$,$Class_3$,…$Class_m$ any tuple S will be member of class with highest posterior probability when condition is S. Meaning that, S is member of $Class_i$ if and only if

   Probability ($Class_i$|S) > Probability ($Class_j$|S) for $1 \leq j \leq$m, j ≠ i

   From Bayes Theorem

   Probability ($Class_i$|S) =

   $$\frac{Probability\ (S|Class_i)Probabilty(Class_i)}{Probability(S)}$$

   a. As Probability(S) has same value for all the classes, only $Probability(S|Class_i) * Probability(Class_i)$ is needed to be maximized.

   b. It is assumed that all attributes are conditionally independent of each other. So, we can rewrite

   Probability (S|$Class_i$) =

   $\prod_{k=1}^{n} Probability\ (s_k|Class_i)$

   = P$robability\ (s_1|Class_i)$ * $Probability\ (s_1|Class_i)$ * P$robability(s_n|Class_i)$

   To calculate Probability (S|$Class_i$) we consider

   i. If $Attribute_k$, is categorical then possibility of $s_k$conditioned on $Class_i$ will be Class $Class_i$'s tuples no with ratio of s_k and $Class_{i,Dt}$ ($total\ tupple\ of\ class\ i\ in\ full\ dataset$) for $Attribute_k$

   ii. $Attribute_k$ is continuous value then we will find out μ$Class_i$and $^{\delta}Class_i$ where $^{\delta}$ is standard deviation and μ is average of attributes of class i

3. To make possibility of class label of S, Probability(S| $Class_i$)*Probability($Class_i$). Then class label of S is considered with highest value of Probability(S|$Class_i$)*Probability($Class_i$).

In this work, at testing phase depending on the features from feature vector, stages of lung cancer will be predicted.

3.3. Support Vector Machine

Support Vector Machine (SVM) algorithm is generally used for the binary categorization purpose creating optimal hyperplane so that the border between the classes becomes maximum. Samples used for building a hyperplane; known as support vectors. An SVM plots input dataset from lower to higher dimensional space to separate dataset efficiently. To classify more than two classes (M-SVM) is used after calculating dual, dual is plotted to high ranking attribute space. To do so kernel trick is applied. Calculating dot product of kernel, discriminant function is prepared without input dataset mapping. Dual is calculated from variables and constraints and not dependent on attribute space.

For this work,

Type of SVM: C-SVC Kernel: RBF

Gamma: 0.71 C: 1

3.4. Decision Tree

C4.5 is one of the most popular algorithms to generate decision tree. Actually, it is extended version of ID3 algorithm. To solve over fitting problem when sample size is small C4.5 was introduced. To select a feature for classifying dataset, Gain Ratio is used.

Gain Ratio of any attribute A = $\frac{Gain\ of\ A}{Spltinfo\ of\ D\ when\ Attribute\ is\ A}$

Gain of A = Information of Dataset - Information of Dataset when Attribute is A.

Information of Dataset = - ∑ Probability of any tuple i * log_2 (Probability of any tuple i)

Split information of D when attribute is

A = - $\sum_{i=1}^{x} \frac{i\ division\ of\ Dataset}{Full\ Dataset} * log_2 (\frac{i\ division\ of\ Dataset}{Full\ Dataset})$

Unpruned: Pruning is performed.

Confidence Factor: 0.50
Number of Folds: 5
Number of Decimal Places: 2
Batch Size: 10
Split Point is not making Actual Value: True
The minimum number of instances per leaf: 3
MDL correction is used when finding splits on numeric attributes: True

3.5. Evaluation

To compare classification results of three algorithms, we have considered some measures like Accuracy, Precision, Recall, F score with the help of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).The main target is to increase both TP and TN while we want to decrease the FP and FN rate. Especially when any disease is predicted FN is needed to be dropped as much as possible. Precision is used to find out Correctly classified Positive sample rate from all positive prediction where

Precision = $TP/_{TP + FP}$

While Recall is measurement of Correctly classified Positive sample rate from all positive samples where

Recall = $TP/_{TP + FN}$

We can consider F score as harmonic mean of precision and recall. As there is tradeoff between Precision and Recall we need to keep track of balance between them using another measure named F score or F-measure

F Score = 2(Recall * Precision) / (Recall + Precision)*

Higher value is expected for F score because higher the value the more perfect relation between Precision and Recall. Maximum is 1 and minimum is 0.

Area under ROC curve has also been considered. It is another measure to compare machine learning algorithm. Value between .9 to 1 is supposed to be very good, .8 to .9 is good and gradually .5 to .6 is not expected value for any types of classifier at all.

3.6. Simulation Tool

We have used here WEKA 3.9, well known free machine learning software. ARFF (Attribute Relation File Format) is a file format of our training and testing data set. This software is also useful to pre-process data set like noise removal, converting nominal values into numeric and so on.

4. Result and Analysis

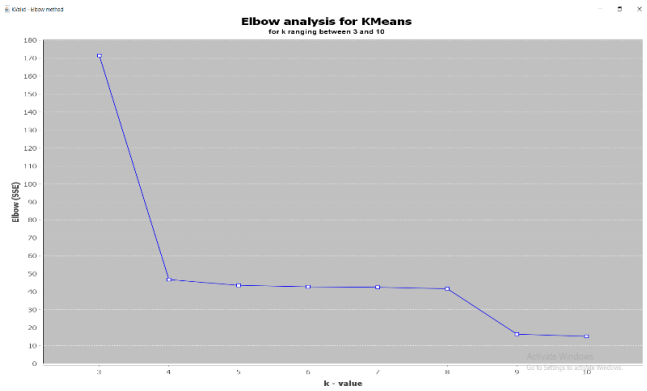

Fig.3.*K-validation for Cluster Number Analysis.*

Here we have compared three most popular machine learning algorithms for our dataset. From figure 4 & 5 we have noticed that for both training and testing phase when we have increased our no. of data samples, accuracy has also been increased in an overall. So, the relation between data samples and accuracy is proportional to one another.

Table 4 Accuracy Rate in Training Phase

Sl.	No. of Test Sample	Accuracy (In percent)		
		Support Vector Machine	Naïve Bayes	Decision Tree
1	317	94.63	98.12	100
2	254	93.7	98.82	100
3	190	92.12	98.95	99.47

Table 5 Accuracy Rate in Testing Phase

Sl.	No. of Test Sample	Accuracy (In percent)		
		Support Vector Machine	Naïve Bayes	Decision Tree
1	140	97.14	100	98.57
2	112	95.24	100	98.12
3	84	89.28	98.82	97.62

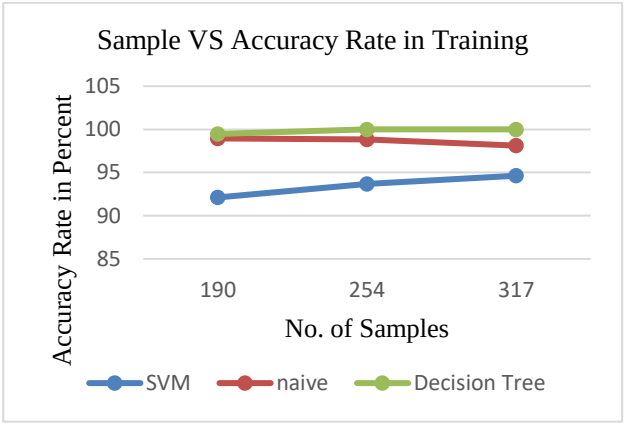

Fig.4.*Sample Vs Accuracy Curve for Training Phase.*

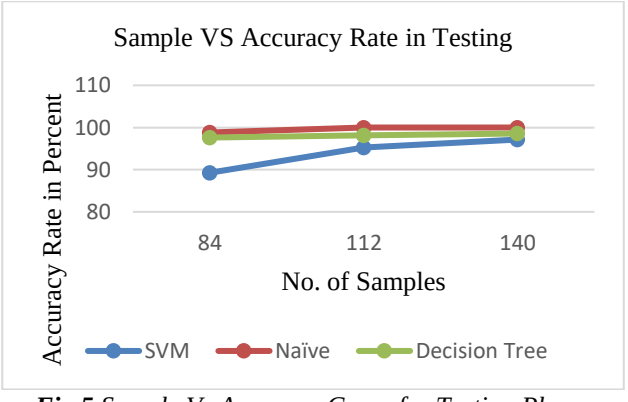

Fig.5.*Sample Vs Accuracy Curve for Testing Phase.*

From Fig.5 it is found that the Naïve Bayes algorithm has the highest accuracy and SVM shows the worst result scenario, though three of them have accuracy over 90%. It was expected that SVM would perform better, but as it is

observed that, data distribution is an imbalanced one and as SVM classifies dataset by constructing linear (and nonlinear) hyper-plane it works worse when data is not balanced.
Now when compare the accuracy of C4.5 Decision Tree and Naïve Bayes algorithm, it's looked out that though accuracy of C4.5 Decision Tree is highest in training phase it has lost the combat to Naïve Bayes in testing phase. This problem can be addressed of over fitting problem meaning that when any network performs so fine for a fixed dataset that can't compile new one further. From Table 2, our sample dataset, we know that samples are continuous not discrete. As a result, to the contrary, when samples are discrete and limited in number Naïve Bayes works better so it's known as continuous classifier. In this network the classification features are defined by hand and there is no pruning. Pruning backfires when we implement decision tree in medical data processing, like our cancer stage detection. Because in C4.5, any sub tree which contains less information gets omitted. But when we are classifying lung cancer stages using Naïve Bayes all information are available for stage grouping purposes. We also considered that our attributes are independent of each other, where Naïve Bayes works best.

Table 5 Various Measurements in Training Phase

Support Vector Machine						
No. of Test Sample	TP	FP	Precision	Recall	F-measure	Area under ROC
317	.946	.031	.947	.946	.946	.958
254	.937	.026	.937	.937	.937	.994
190	.921	.039	.931	.921	.916	.941
Naïve Bayes						
No. of Test Sample	TP	FP	Precision	Recall	F-measure	Area under ROC
317	.981	.008	.982	.981	.981	1
254	.988	.005	.989	.988	.988	1
190	.989	.007	.990	.989	.989	1
Decision Tree						
No. of Test Sample	TP	FP	Precision	Recall	F-measure	Area under ROC
317	1	0	1	1	1	1
254	1	0	1	1	1	1
190	.995	.001	.995	.995	.995	.99

Table 6 Various Measurements in Testing Phase

Support Vector Machine						
No. of Test Sample	TP	FP	Precision	Recall	F-measure	Area under ROC
140	.971	.004	.977	.971	.972	.984
112	.952	.027	.972	.952	1	.962
84	.893	.049	.948	.893	1	.922
Naïve Bayes						
No. of Test Sample	TP	FP	Precision	Recall	F-measure	Area under ROC
140	1	0	1	1	1	1
112	1	0	1	1	1	1
84	.988	.005	.989	.988	.988	1
Decision Tree						
No. of Test Sample	TP	FP	Precision	Recall	F-measure	Area under ROC
140	.986	.002	.987	.986	.986	.998
112	.991	.001	.992	.991	.991	.99
84	.976	.004	.980	.976	.976	.995

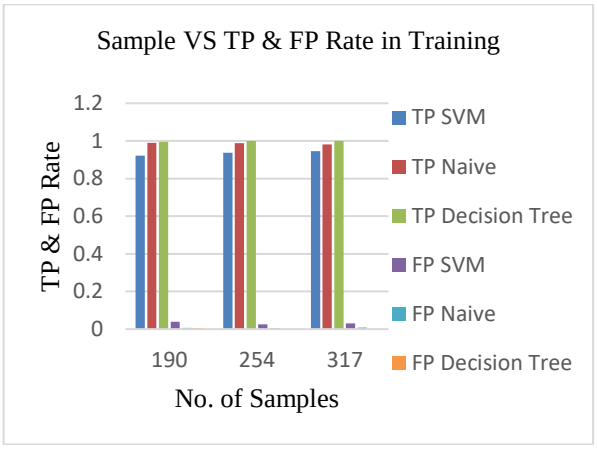

Fig.6.*Sample Vs TP, FP rate in Training Phase.*

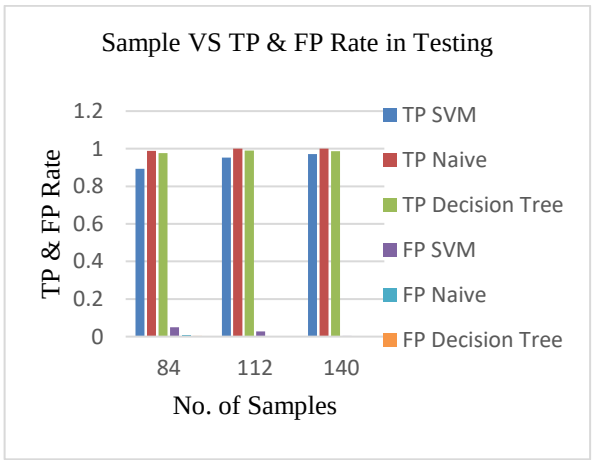

Fig.7.*Sample Vs TP, FP rate in Testing Phase.*

From figure 6 and 7, for all three classifiers, TP rates are higher than FP which was expected and in testing phase as the best classifier, Naïve Bayes Classifier has better TP rate than other two algorithms. From figure 8 and 9, it is observed that, in the training phase Decision Tree has higher Precision and Recall rates while in the testing phase naïve Bayes classifier has the best rate.

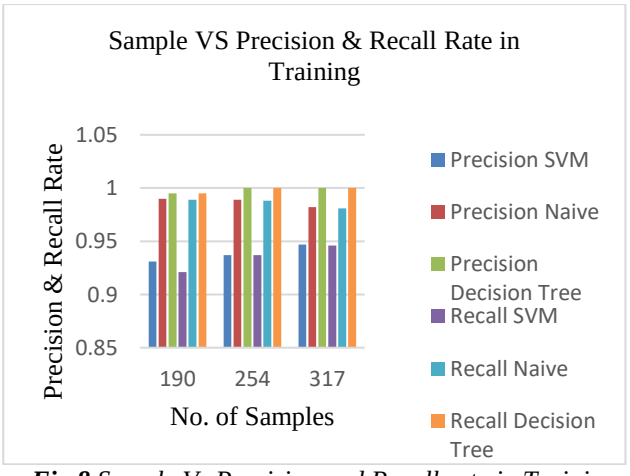

Fig.8.*Sample Vs Precision and Recall rate in Training phase.*

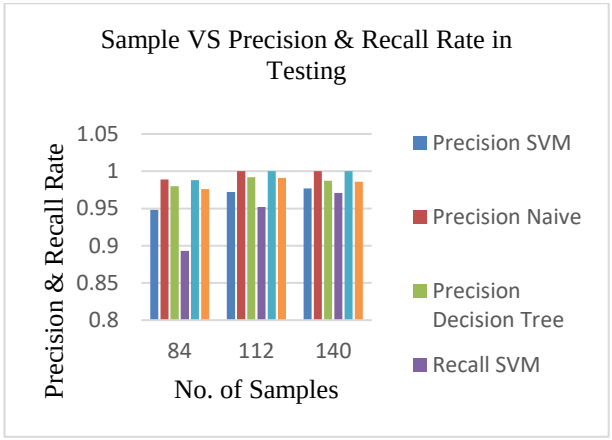

Fig.9.*Sample Vs Precision and Recall rate in Testing phase.*

As there is always tradeoff between Precision and Recall, F-value is calculated and for both training and testing phase F-value is tends to 1 on average which is considered as best value for F-value or F-measure

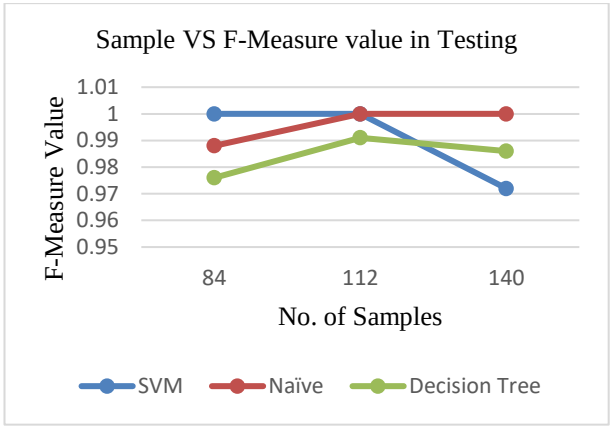

Fig.10.*Sample Vs F- Value in Training phase.*

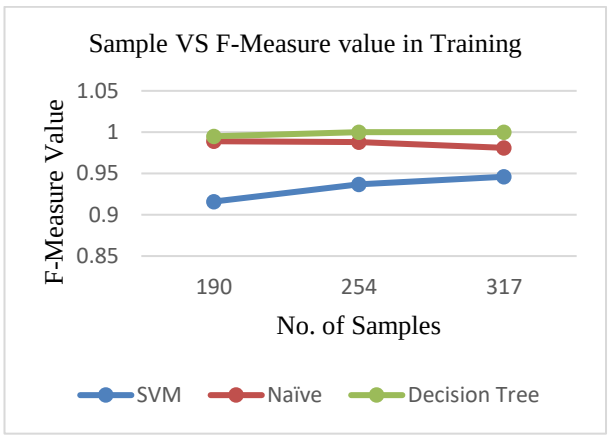

Fig.11.*Sample Vs F- Value in Testing phase.*

Area value under ROC curve is in a satisfactory level both in training and testing phase. Average value of both Naïve Bayes algorithm and Decision tree is competent. But with increased accuracy in the testing phase values of Naïve Bayes are 1.

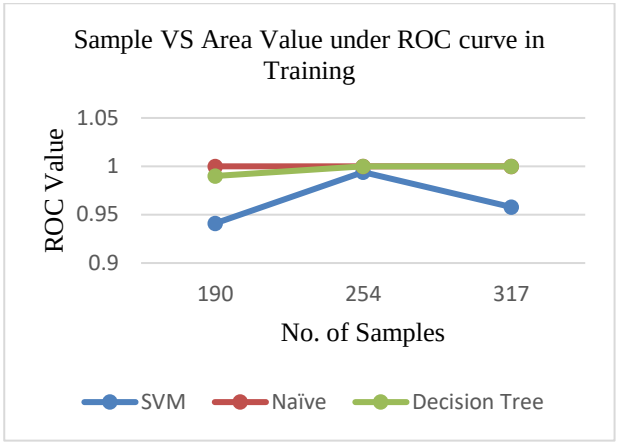

Fig.12.*Sample VS Area Value under ROC curve in Training Phase.*

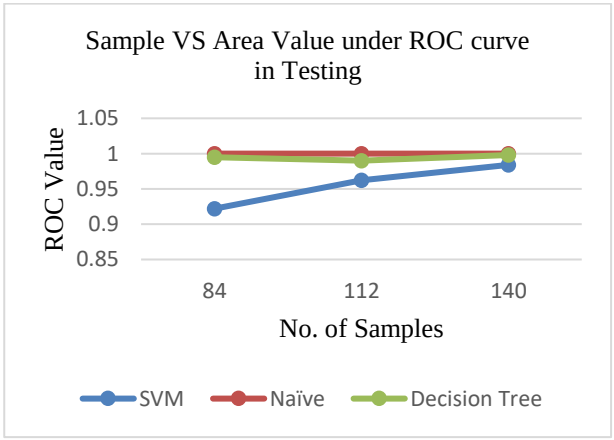

Fig.13.*Sample VS Area Value under ROC curve in Testing Phase.*

5. Conclusion

Naïve Bayes is always considered as a better classifier than others for computer vision and medical data processing. In this work it has performed best again among three most common classifier networks. The human doctor usually predicts cancer stages after analysing CT scan images, now this job can be completed with average 99.6% using our proposed network following Naïve Bayes Algorithm as an automated manner. Not only accuracy, using different parameters like F-measure, ROC curve calculation it is found better than two others. As Decision Tree has also shown considerable result quite similar as Naïve Bayes in different parameters, it is our second-best algorithm. To pre-process the dataset that is image processing step helped to calculate our features in a delicate details format. As there were difficulties to collect real life raw images, our dataset was not rich enough. Surpassing obstacles four features have been picked up from processed images. If we can build rich data set by stretching more characteristics and if we try different feature combinations, it will increase our model validity and can be possible to obtain 100% accuracy. As automated methods to predict and detect various diseases in medical science are becoming vogue day by day, it is hoped to implement our proposed work to detect lung cancer stages in a self-acting system.

6. References

[1] Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol*. 2017;7(9):170070. doi:10.1098/rsob.170070

[2] Cancer survival in England: adult, stage at diagnosis and childhood patients followed up to 2018," *Home - Office for National Statistics*. [Online]. Available: https://www.ons.gov.uk/.

[3] Carson J, Finley DJ. Lung cancer staging: an overview of the new staging system and implications for radiographic clinical staging. *SeminRoentgenol*. 2011; 46(3):187–193. doi:10.1053/j.ro.2011.02.004

[4] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L. and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), pp.394-424

[5] M. A.gajdhane and P. D. L.m, "Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques," IOSR Journal of Computer Engineering, vol. 16, no. 5, pp. 28–35, 2014

[6]Implementation of Lung Cancer Nodule Feature Extraction Using Digital Image Processing KHIN MYA MYA TUN1, AUNG SOE KHAING2

[7] A.Amutha and R.S.D Wahidabanu, "A Novel Method for Lung Tumor Diagnosis and Segmentation using Level SetActive Contour Modelling", European Journal of Scientific Research, Vol.90, No.2, November 2012, pp.175-187

[8] Mokhled S. Al-Tarawneh, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, Issue 20, JanuaryJune 2012, p.147-158. ISSN 1583-1078

[9] N.A. Memon et. al, "Segmentation of Lungs from CT Scan Images for Early Diagnosis of Lung Cancer," World Academy of Science, Engineering and Technology. 2006

[10] M.-S. Brown, M.-F. McNitt-Gray and N.-J. Mankovich, Method for segmenting chest CT image data using an anatomical model: preliminary results, IEEE Trans. on Med. Imaging, vol. 16, No. 6, pp. 828839, 1997.

[11] C.-W. Bong, H.-Y. Lam and H. Kamarulzaman, A Novel Image Segmentation Technique for Lung Computed Tomography Images, Communications in Computer and Information Science, Vol. 295, pp. 103-112, 2012.

[12] J.Wang, R. Engelmann and Q. Li, Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral scanning technique, Medical Physics, Vol. 34, No. 12, pp. 46784689, 2007

[13] D.-T. Lin, C.-R. Yan and, W.-T. Chen, Autonomous detection of pulmonary nodules on CT images with a neural network-based fuzzy system, Comput. Med. Imaging Graph., Vol. 29, pp. 447-458, 2005

[14] M. Antonelli, M. Cococcioni, B. Lazzerini and F. Marcelloni, Computer-aided detection of lung nodules based on decision fusion techniques, Pattern. Anal. Applic., Vo. 14, pp. 295310, 2011.]

[15]J.Wang, R. Engelmann and Q. Li, Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral scanning technique, Medical Physics, Vol. 34, No. 12, pp. 46784689, 2007.

[16]D.-T. Lin and C.-R. Yan, Lung nodules identification rules extraction with neural fuzzy network, In Proc. of the 9th IEEE International Conference of Information Processing (ICONIP), Vol. 4, pp. 2049-2053, Singapore, 18-22 November 2002

[17] PMC, E., 2020. *Europe PMC*. [online] Europepmc.org. Available at:<http://europepmc.org/articles/PMC3351680>

[18] Dhara, A., Mukhopadhyay, S., Dutta, A., Garg, M. and Khandelwal, N., 2020. A Combination Of Shape And Texture Features For Classification Of Pulmonary Nodules In Lung CT Images.